基于词向量扩展的学术资源语义检索技术*

王仁武 陈川宝 孟现茹
华东师范大学经济与管理学部信息管理系 上海 200241

摘要:[**目的/意义**]尝试以统计的方法为指导思想,探究基于词向量扩展的语义检索技术来提升学术资源的语义检索能力。[**方法/过程**]利用自然语言处理、文本挖掘技术,对采集来的学术资源(主要是学术论文)元数据进行预处理,结合 word2vec 词向量生成工具和 elasticsearch 全文检索引擎搭建语义检索系统,对学术资源进行语义检索的探索研究。[**结果/结论**]本文提出的方法能够有效提升学术信息的检索效果,一定程度上实现学术资源的语义检索,并为后续语义检索的进一步研究提供借鉴。

关键词: word2vec Elasticsearch 语义检索 学术资源
分类号: G250
DOI:10.13266/j.issn.0252-3116.2018.19.014

1 引言

学术用户的信息需求往往具有专业性、知识性、个性化、多元化、便捷性和人性化的特点[1]。为了满足这类用户特殊的信息需求,相应地出现了许多学术资源数据库:Web of Science、Science direct、中国知网、万方学术等。这些学术数据库拥有海量的学术资源,一直以来都在为用户提供着高质量的学术信息服务。但是种类繁多的学术资源库也给用户的使用带来了一定的麻烦:用户往往需要在几个类型的学术数据库之间进行查询,才能全面、准确地获得自己所需要的信息;同时,学术论文检索数据库界面复杂,检索式繁多,对于新手来说人机交互效果很不友好;而且,大多数的学术资源数据库目前主要还是基于关键词匹配的检索,用户在信息检索的时候会出现"词汇问题"[2],导致信息检索系统在面对同样的检索意图时,因为用户给出的关键词不同而返回出截然不同的检索结果,最终影响检索质量。

如果在进行学术信息搜索的时候,相关学术检索系统能够提供一个简洁的搜索框,并且能够突破用户给出的关键词局限,理解用户真正检索意图,实现语义层面的信息检索,无疑将大大提升学术信息检索的效果。本文尝试研究利用基于深度学习的词向量技术,通过语义扩展来提升学术信息的检索效果,一定程度上实现学术资源的语义检索。

2 相关研究

2.1 语义检索研究

目前实现语义检索的方法大致可以分为基于规则和基于统计两类:基于规则的语义检索主要通过人工撰写规则知识库的方式,构建一个语义知识网,在其中进行语义推理;基于统计的语义检索以数理统计为工具,不要求在事前对相应知识规则进行人工构建,而是通过相应算法在大规模语料库中进行总结,归纳出词和文档之间的统计信息,随后对词语进行语义相似度的计算,并用于语义检索。

语义知识库是基于规则的语义检索较早的一种实现方式,目前较为著名的两个语义知识库是 WordNet 和 HowNet(知网)。D. I. Moldovan 和 R. Mihalcea[3]在对查询语句进行相关处理后,使用 WordNet 中的词汇对查询请求进行查询词扩展,定义其查询词的同义词集合,并应用到 AltaVista 检索系统中去。高雪霞与炎士涛[4]提出一种基于 Jaccard 系数的词义消歧方法,以 WordNet 词库为基础,对查询词中的歧义词进行消歧,在检索结果的精确度方面较以往的信息检索系统提高了 10%。王李冬与张慧熙[5]以国内新浪微博平

* 本文系国家社会科学资金项目"基于数据驱动的图书馆资源发现平台研究"(项目编号:16BTQ026)研究成果之一。
作者简介:王仁武(ORCID:0000-0003-0980-0779),副教授,博士,硕士生导师,E-mail:rwwang@infor.ecnu.edu.cn;陈川宝(ORCID:0000-0003-4121-2328),硕士研究生;孟现茹(ORCID:0000-0003-2899-6471),硕士研究生。
收稿日期:2018-04-09 修回日期:2018-06-14 本文起止页码:111-119 本文责任编辑:易飞

台的文本为研究对象，基于 HowNet 按照语义相关度对中文待检索的主题词和新浪微博的文本词汇进行匹配，以满足用户的查询意图，对微博的短文本语义检索进行了尝试。

基于统计的方法一直都以其自身严谨、科学的特点，成为人们在实现语义检索时的首选方法。近年来随着芯片技术和机器学习算法的发展，给计算机带来了更强劲的算力和更强大的语义理解能力，计算机对统计方法的支持，使得统计方法凭借其高效、快捷的特点，在语义检索研究方面又焕发新春。2003 年 D. M. Blei 等[6-7]提出的 LDA 主题模型使得人们得以从主题相关词的概率统计的角度实现语义检索。刘启华[8]随后基于 LDA 主题模型设计了 PMM 模型和 TBS 模型，实现的语义检索系统能够有效提升系统检索效果。Google 在 2013 年推出了 word2vec[9][10][11]词向量生成工具，使得人们能够从大规模的文本语料中进行词向量的训练，得到高质量的词向量以应用到后续的自然语言处理任务中去。范桥青和方钰[12]以 Reuters－21578 和 120ask 中的文本为语料库，利用 word2vec 训练出的词向量来比较词语间的语义相似度；并将训练好的词向量结合 Axiomatic 最优检索模型，实现面向健康问答社区的语义检索。刘梦兰等[13]同样以 word2vec 为词向量训练工具，结合专利文献自身的特点，提出了一种基于词向量的查询扩展方法，有效提升专利文献的检索效果。许稳堂[14]以微博文本为研究对象，利用 word2vec 中的 skip-gram 模型训练微博文本的词向量，通过词向量的加权平均以获得微博文档及查询语句的向量表示，设计并实现了一种基于 MRA-E 算法的微博语义检索系统。Word2vec 之后，斯坦福大学也开放了一种基于全局的词向量训练工具 Glove[15]，陈国华等[16]基于 Glove 训练词向量，利用随机映射的方法，在大规模的向量空间中快速定位向量，并提出了一种学术文档向量化的方案，在随后的学者网学术检索中取得良好的检索效果。2018 年，Google 更是基于词向量技术，向外界推出了 AI 检索引擎 semantic experiences[17]，用户可以使用自然语言与检索系统对话，系统根据用户的提问，而不是拘泥于关键词，回答用户问题。

2.2 word2vec

本文采用 word2vec 生成工具来训练学术文本的词向量。word2vec 是 Google 在 2013 年向外界推出的一款词向量生成工具，由 T. Mikolov 领导的研究小组研发。word2vec 将文本语料库作为输入，利用其内部的两个神经网络语言模型从语料库中学习词汇的向量表示，再以词向量作为输出。

word2vec 中有两种训练词向量的框架（见图 1），用以预测：①给定上下文 的情况下，词 w 的概率 的 CBOW（continuous bag-of-words）；②给定词 w 的情况下，其上下文 的概率 的 Skip-gram（continuous skip-gram）。在训练过程中，两种架构又各有侧重：CBOW 在词向量的训练速度方面表现出色；Skip-gram 虽然在训练速度上较慢，但是其训练低频词的效果较好。

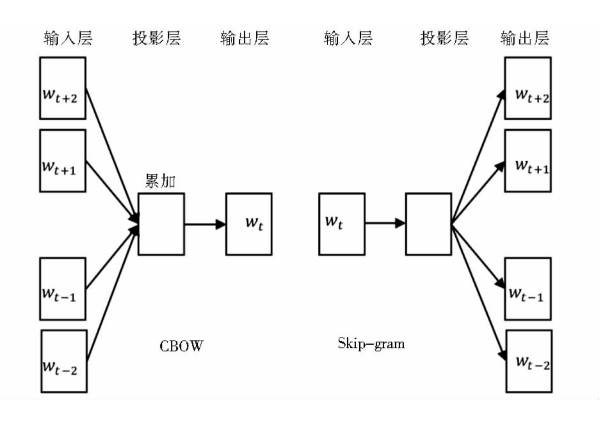


图 1 word2vec 的 CBOW 和 Skip-gram 框架

3 语义检索系统架构

本文设计了如图 2 所示的一种基于词向量技术的学术资源语义检索系统。其设计思路是从对现有的语义检索模型和系统的分析入手，将基于深度学习的词向量的文本语义处理技术与开源的全文搜索引擎 elasticsearch 相结合融入到语义检索中，建立基于词向量扩展的语义检索系统模型；然后将该模型应用于学术资源的语义检索服务领域，并对其语义检索效果进行分析与评估。

该系统主要有 5 个组成部分：数据搜集与处理模块、词向量模块、查询扩展模块、全文检索模块和数据分析模块。具体结构见图 2。

数据搜集与处理模块主要负责的是整个语义检索系统所需的学术文献资料数据的搜集任务，可以用人工导入、网络爬虫抓取、调用接口读取的方式从个人信息资源库、专业数据库和互联网等渠道对语义检索系统所需的文献数据进行搜集，并对收集来的数据进行数据质量检查以及数据清洗等工作。最终准备好文档对象和规范化的数据，提交给全文检索模块和词向量模块进行文档的索引和词向量的训练。

词向量模块主要负责训练出词语的词向量以对用户查询进行词向量语义查询扩展的应用。同时词向量

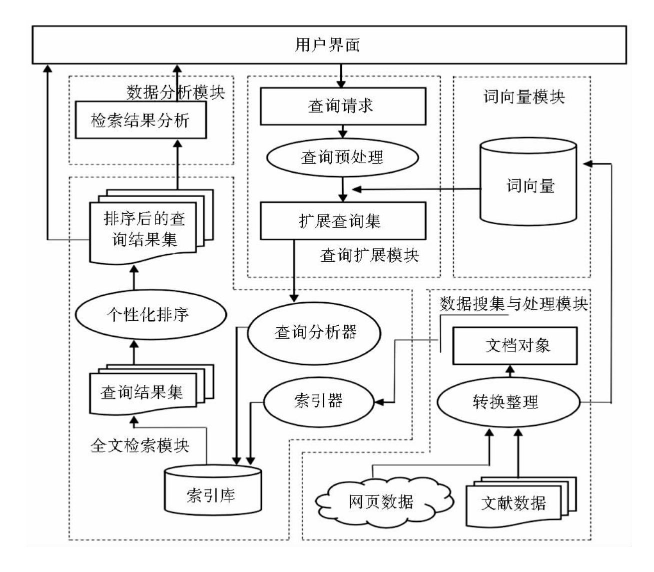


图2　学术资源语义检索系统的组成结构

库也会随着数据集的更新而不断进行更新，确保其中词语的实效性和覆盖率。

查询词扩展模块主要负责用户查询请求的查询词扩展工作。首先是接受用户的查询请求，对用户输入的查询请求进行分词、去停用词等操作，将处理规范的查询请求利用词向量模块中的词向量库对检索词进行扩展，并将扩展查询集提交给后续的全文检索模块进行用户查询请求的全文检索活动。

全文检索模块主要负责具体的全文检索任务。首先是将数据搜集与处理模块整理好的文档对象导入全文检索引擎中，依据相应的分析器进行文档的索引操作，形成文档的索引库。其次是依据传进来的查询请求，在索引库检索符合要求的文档，形成初检结果。最后依据特定的相关性评分改进算法对初检结果进行相关性评分改进，使之能够更加符合用户的需求，按照一定的排序提交到用户界面，将结果展示给用户。

数据分析模块负责的是对检索结果进行数据分析和数据可视化的任务。在得到检索结果的同时，用户可能还想知道这些检索结果内在的一些数据关联。对检索结果进行数据分析操作，并将分析结果以可视化图表的形式返回给用户，能够帮助用户发现检索结果的内在规律并从整体把握检索结果特征，以快速定位自己所需文献，甚至启发用户思考，进一步厘清自身信息需求，从而进行二次检索。

4　系统关键技术研究

4.1　领域词典构建技术

所谓领域词典是指一种记载了特定研究领域的专有词汇或术语的词典。分词工具内置的通用词典对特定领域内的专有名词或术语收录不足，这会使得分词工具错误切分相关专有名词或术语，例如，将“支持向量机”切分成“支持/向量/机”，“潜在语义分析”切分成“潜在/语义/分析”，“布拉德福定律”切分成“布/拉德福/定律”，等等。

我们首先利用采集来的比较规范的学术资源元数据，“Author-作者”“Keyword-关键词”和“Organ-机构”字段中的数据比较规范，可以直接引进作为领域词。为尽可能多地获取领域词语，本文又以点互信息和词频为统计标准，从语料库中继续抽取领域词。

点互信息(piontwise mutual information，以下统一简称PMI)能够刻画两个随机变量之间的关联程度，在领域词的识别任务中，可以用来衡量字符间的结合程度。其计算公式如式(1)所示：

$$PMI(x,y)=\log\frac{P(x,y)}{P(x)P(y)} \qquad 式(1)$$

其中，PMI(x,y)就是字符串x,y的点互信息值，也就是它们的相关程度；P(x)表示字符串x出现的概率，P(y)表示字符串y出现的概率，P(x,y)表示字符串x,y共现的概率。当PMI值小于或者等于零时，说明二者没有明显的关联性；当PMI值大于零时，说明二者相关关系较强，PMI取值越大，二者的相关关系也就越强。

我们定义分词后得到的每个独立字符组合，无论该组合内包含多少个字符，都称之为“词单元”。例如，一串字符 $s_1s_2s_3s_4s_5s_6$，经过分词后得到 $s_1s_2 | s_3 | s_4s_5s_6$，则 s_1s_2、s_3、$s_4s_5s_6$ 都分别为一个词单元。而连续的词单元组成一种模式，有几个词单元，我们就称该模式为几维模式，如 $s_1s_2 | s_3 | s_4s_5s_6$ 为三维模式。对于领域词典选出来的候选词 $s_1s_2 | s_3 | s_4s_5s_6$，$s_1s_2 | s_3$ 和 $s_3 | s_4s_5s_6$ 来说，$s_1s_2 | s_3$ 和 $s_3 | s_4s_5s_6$ 都是候选词模式 $s_1s_2 | s_3 | s_4s_5s_6$ 的子模式。

图书情报领域的专业术语一般稳定在10个字以内，也就是经常会被4个词单元内的组合所覆盖到。另外，张榕[18]在术语数据库中的统计研究也表明，由2－4个词语组合而成的术语占到总体的71.723%。因此，在这里我们只考虑2个、3个和4个词单元组合出来的候选词情况，也就是对2维、3维和4维候选词进行抽取。

领域词典候选词的选取步骤为：

(1)语料库预处理。也就是去除相关乱码，并以空格替代原先语料库中的中英文标点符号、数字。随后以哈尔滨工业大学停用词表对语料中的停用词进行

处理,将文中的停用词也替换成空格。

(2)分词并统计词单元词频。使用 python 编程,利用 jieba 分词模块对语料进行分词,得到分词后的语料库,并对其中分出来的每一个词单元进行词频统计,为后续的 PMI 计算做好准备。

(3)识别候选词。以经过分词的语料作为遍历素材,以第一个词单元为开始进行循环遍历,分别考察循环到的当前词单元的后面第一个、第二个和第三个词单元是否为空,如果不是,分别将它们记录进 2 维、3 维和 4 维候选词列表,如果其中任何一个词为空,则立即终止后续的判断,进入下一循环。遍历结束后,我们就能得到所有可能的 2 维、3 维和 4 维候选词列表。

(4)候选词初次筛选。经过上述步骤得出来的候选词只是词单元的简单组合情况,并不能当作一种领域词汇来用。以 2 维候选词开始,首先对各个候选词及组成该候选词的词单元进行词频统计,在此基础上再计算该候选词的 PMI 值。对于不符合预先设定的词频和 PMI 阈值的候选词予以删除处理,剩下的词汇即为初次筛选合法的 2 维候选词(在这里,合法意为可以认定为该词汇是正常的词汇的意思)。再利用这些合法的 2 维候选词去筛选 3 维候选词。即,合法的 3 维候选词应当是合法的 2 维候选词的扩展形式,删除那些不包含合法 2 维候选词的 3 维候选词,并在此基础上再计算那些剩下的 3 维候选词的词频和 PMI 值,并依据词频和 PMI 阈值进行筛选,得到初次筛选合法的 3 维候选词。最后依据合法的 3 维候选词对 4 维候选词进行筛选,做法如利用 2 维候选词对 3 维候选词进行筛选的过程一样,最后得到初次筛选合法的 4 维候选词。

其中,阈值的选取对候选词的确定有着很大的关系。经多次试验,PMI 和词频阈值定在 5 和 20 得到的候选词数量和质量都较为合理。

在抽取出领域词的候选词集之后,以如下的规则对候选词进行筛选和剔除:删除表示时间日期和表示数量的词汇;合并某种模式的子模式;删除"该""应""这"等单字辅助词开头或者结尾的字符串;删除"非常""十分""很难"等包含表示程度的词汇。

4.2 词向量语义扩展技术

(1)语义概念扩展词来源。语义概念扩展所需的词语来源是学术资源元数据。在对采集来的学术资源元数据进行筛选、去除标点符号、分词及去除停用词等处理之后,将其作为语料导入 word2vec 进行词向量的训练,最终得到该语料库中词语的向量表示及各个向量之间的语义距离。

(2)语义概念扩展词选取标准。语义距离方面。word2vec 能够根输入词返回其语义相近的词及其语义相似度。语义相似度越接近于 1,则说明两个词汇的语义越相近。在这里我们选取的语义相似度阈值是 0.85,当两者之间的语义相似度大于这个阈值时,则认为两者存在较强的语义关联,可以将该词汇作为查询扩展词加入检索式中去。

语义扩展词的最大个数方面。查询扩展本意是通过多提供额外信息的方式,帮助检索系统明确用户的检索意图,从而实现更好的检索效果。但是扩展的查询词过多也会带来额外的信息干扰。在进行查询词扩展的时候,需要把握好检索词数量和检索结果集数量之间的关系。本文采取与 word2vec 默认返回最相似的 10 个词汇相同的标准,也就是说,最多每个原查询词会有 10 个符合语义距离阈值的查询扩展词。

(3)基于词向量的语义查询扩展流程。本文基于词向量的语义查询扩展流程分为以下几个步骤:首先,用户进行检索查询输入。然后对用户的查询输入进行查询预处理,预处理操作包括利用导入了领域词典的 jieba 分词算法进行分词、利用哈工大停用词表去除停用词、去除标点符号等。之后利用训练好的词向量库对用户查询进行扩展,对于符合语义关系阈值的词汇,加入到查询扩展词中去,与原查询词形成查询扩展集合。最后,将查询扩展集合输入检索系统,进行信息检索。

4.3 针对学术文献的个性化评分方案

本文选取 Elasticsearch[19] 作为语义检索系统的检索引擎,Elasticsearch 默认使用的是 Lucene 的内部的 TF/IDF(词频/反文档)文档相关性评分机制,其计算公式如式(2)所示:

$$\mathrm{score}(q,d) = queryNorm(q) * coord(q,d) * \sum_{t\ in\ q}(tf(t\ in\ d) * idf(t)^2 * t.getBoost() * norm(t,d)) \quad 式(2)$$

式(2)的各个组成因子释义见表 1。

上述 Elasticsearch 的文本相关性评分算法在实际应用中已经有着很好的表现,但是在针对学术文献进行检索时,其表现力还是略有不足,主要体现在:①其未能考虑查询词出现在文献中的位置所带来的相关性影响;②其未能考虑原查询词和扩展查询词之间的语义差别;③其未能考虑文献被引情况所体现的文献价值。

表1 Elasticsearch默认评分公式中的组成因子及其含义

编号	组成因子	因子含义
1	score(q,d)	被检索文档 查询请求 的吻合程度,也就是文本的相关度
2	queryNorm(q)	查询归一化因子。在不影响文本相关性排名的情况下,归一化因子对文本的相关性评分进行归一化处理,使得最终的文本相关性得分稳定在一个区间内,方便不同查询方式的相关度分数的比较。归一化通常采用的计算方法是: $queryNorm(q)=\frac{1}{\sqrt{sumOfSquaredWeights}}$ 其中,sumOfSquaredWeights是查询请求q中词项t的idf平方和
3	coord(q,d)	协调因子。用来刻画检索请求q中的词项t被文档匹配到的比例,越多的查询项在一个文档中,说明文档的匹配程度越高。协调因子通常采用的计算方式是: $coord(q,d)=\frac{overlap}{maxoverlap}$ 其中,overlap表示文档d包含的检索请求q中词项t的个数,maxoverlap表示检索请求q中的全部词项t个数。
4	tf(t in d)	词频因子。用来刻画检索请求q中的词项t在被检文档d中出现的次数,出现的次数越多,被检索文本与检索请求的相关性就越高。词频因子通常采用的计算方法是: $tf(t\ in\ d)=\sqrt{frequency}$ 其中,frequency表示检索请求q中的词项t在被检文档d中出现的次数
5	idf(t)	逆词频因子。用来刻画检索请求q中的词项t的在所有文档中出现的逆文档频率,出现频率越高,说明该词汇特殊性不是很强,评分权重就越低。反之,说明该词汇很具有特殊性,能够区分出文档的性质,评分权重越高。逆词频因子通常的计算方法是: $tf(t\ in\ d)=1+\log\frac{numDocs}{docFreq+1}$ 其中,numDocs为文档总数,docFreq为词项t在所有文档中出现的次数
6	getBoost()	预设定的权重因子。在为文档建立索引时,为每个字段所预先设定的权重值,默认情况下是1
7	notm(t,d)	标准化因子。是在索引时为字段所分配的权重值与字段长度的归一之和

关于上述的第一和第二点不足,张孝飞和孔敏秀[20]建议学术文献中文献的各个分块(即“题名”“关键词”“摘要”和“正文”等字段)和扩展查询词应该在文献检索中分配不同的价值权重,并给出了相应权重的取值和计算方式。而对于第三点不足,著名的PageRank[21]算法则给出了一个比较好的思路。即,被引量较高的学术文献其往往具有较高的质量或者较显著的代表性,在文本相关性评分时,其评分应予以适当的提升。

在此,本文提出一种适用于学术文献语义检索的个性化评分策略。首先是学术文献的不同字段权重值方面,经多次试验,对“题名”“关键词”和“摘要”字段,分别赋予1.2、1.1和1的权重,在建立文档的映射的时候以boost值指定。也就是说,其不同字段在相关性评分时的重要程度是:“题名”>“关键词”>“摘要”。然后是原查询词与扩展查询词的权重分配方面,原查询词予以1的权重分配,而扩展查询词的权重取值则是词向量给出的其与原查询词的语义关联值,在查询时以boost值指定。最后是文献被引情况对文本相关性评分的影响方面,利用DSL中function_score函数的field_value_factor参数,将文献的被引用量纳入考量,以log对数函数对被引量citation进行平滑处理,避免过高的被引量带来的干扰。其具体的计算方式是:log(1+citation),在此基础上与原相关性评分进行相加操作(sum),则文本的最终相关性得分如式(3)所示:

最终相关性得分=未考虑文本引用量的相关性得分+log(1+citation)　　式(3)

5 系统实现与评估

5.1 实验源数据来源

此次实验使用的论文元数据采集自中国知网中2002－2017年这15年间的图书馆、情报与档案(以下简称“图情档”)领域核心期刊文献元数据,总计122 519篇文献元数据。同时在中国知网中,以网络爬虫的方式对这122 519篇文献的被引量进行爬取,用于文档相关性评分的加权改进。图情档领域核心期刊选取标准是依据2017年最新版北大核心期刊要目总览中G25图书馆事业、信息事业和G27档案事业栏目中收录的28种核心期刊[22]。已采集的122 519篇论文元数据所含字段及其含义见表2。

5.2 领域词典分词效果评测

根据前面设计的领域词典自动构建算法,我们一共得到初始2维候选词432 457个,3维候选词335 062个,4维候选词157 853个。在对候选词进行PMI值、词频值的甄选和规则过滤之后,最终得到4 393个正式领域词。其中,2维候选词3 590个,3维候选词681个,4维候选词122个。

结合直接引进的223 831个论文源数据相关词汇,

表2 学术资源元数据字段及其含义

编号	元数据字段	字段含义
1	DataType	文献类型
2	Title-题名	文献题名
3	Author-作者	文献作者
4	Source-刊名	文献发表的期刊
5	Year-年	文献发表的年份
6	PubTime-出版时间	文献出版时间
7	Keyword-关键词	文献关键词
8	Summary-摘要	文献摘要
9	Period-期	刊载于期刊第几期
10	Roll	刊载于期刊第几卷
11	PageCount-页数	文献所占页数
12	Page-页码	文献所在页码
13	SrcDatabase-来源数据库	文献来源数据库
14	Organ-机构	发文作者所属机构
15	Link-链接	文献链接
16	Citation	文献被引量

与jieba分词自带的字典中的词汇去重，最后得到205 826个领域词汇加入到后续的分词活动中。

其中，自动抽取的2维、3维和4维的领域词示例如表3－表5所示：

表3 2维领域词部分示例

编号	词语	PMI值	词频
1	城市圈	7.495 660 692	30
2	语义关系	5.515 333 473	230
3	舆情监控	8.148 695 785	45
4	智能终端	8.202 143 507	24
5	情感词典	9.670 090 897	35
6	私有云	9.133 067 131	30
7	数字仓储	5.334 687 452	41
8	决策咨询	5.081 141 5	70
9	时间序列	9.563 574 419	91
10	浅阅读	7.063 534 789	92

表4 3维领域词部分示例

编号	词语	PMI值	词频
1	网络出版总库	8.274 331 619	154
2	印刷型文献	6.866 084 784	41
3	信息生态链	9.352 306 309	440
4	农家书屋工程	6.964 913 295	34
5	跨语言检索	6.858 043 683	22
6	支持向量机	8.029 998 022	89
7	反竞争情报	8.723 888 938	173
8	机构典藏库	8.065 792 649	26
9	命名实体识别	10.962 828 08	38
10	联机合作编目	11.754 596 34	30

表5 4维领域词部分示例

编号	词语	PMI值	词频
1	供给侧结构性改革	11.136 396 89	26
2	文献资源共建共享	9.413 718 519	105
3	模糊综合评判法	6.053 802 481	26
4	联合数字参考咨询	7.057 873 748	26
5	政府信息公开条例	10.743 964 72	213
6	中国科学引文数据库	10.087 290 84	27
7	儿童阅读推广活动	7.608 196 978	26
8	人大复印报刊资料	10.856 380 89	36
9	协同过滤推荐算法	10.433 849 17	30
10	农村公共文化服务	6.615 571 814	47

之后，我们将领域词典导入jiaba分词的用户词典，在语料库中与未添加领域词典的分词效果进行分词对比实验。

未添加领域词典的jieba分词算法的分词效果如图3所示：

针对/图书/借阅/量/数据/呈现/的/非/平稳/动态随机/变化/特性/ /采用/支持/向量/机/作为/建模/工具/ /利用/混沌/时间/序列/理论/对/图书/借阅/流量/行为/进行/了/建模/和/学习/预测/ /结果表明/ /该/方法/可/有效/解决/图书/借阅/行为/中/的/非线性/问题/ /预测/结果/合理/ /对/提高/图书/借阅/管理/质量/有/较/好/的/理论/和/实践/参考价值

基于/品牌/建设/的/视角/ /以/海南/职业/技术/学院/ /演练/说/ /立体/阅读/推广/活动/模式/为例/ /分析/高职/院校/图书馆/立体/阅读/推广/品牌/建设/的/内涵/和/意义/ /总结/高职/院校/图书馆/立体/阅读/推广/活动/存在/的/问题/并/提出/相应/的/对策/ /为/构建/具有/高职/院校/图书馆/特色/的/阅读/推广/品牌/提供/借鉴/和/参考

图3 未添加领域词典的jieba分词算法的分词效果

添加了领域词典的jieba分词算法的分词效果如图4所示：

针对/图书借阅量/数据/呈现/的/非/平稳/动态随机/变化特性/ /采用/支持向量机/作为/建模/工具/ /利用/混沌时间序列/理论/对/图书借阅流量/行为/进行/了/建模/和/学习/预测/ /结果表明/ /该/方法/可/有效/解决/图书借阅/行为/中/的/非线性/问题/ /预测/结果/合理/ /对/提高/图书借阅/管理/质量/有/较/好/的/理论和实践/参考价值

基于/品牌建设/的/视角/ /以/海南职业技术学院/ /演练/说/ /立体/阅读推广活动/模式/为例/ /分析/高职院校图书馆/立体/阅读推广/品牌建设/的/内涵/和/意义/ /总结/高职院校图书馆/立体/阅读推广活动/存在的问题/并/提出/相应/的/对策/ /为/构建/具有/高职院校图书馆/特色/的/阅读推广/品牌/提供/借鉴/和/参考

图4 添加领域词典后的jieba分词算法的分词效果

从这两个分词效果的测评片段我们可以看出，加入领域词典后的分词效果更为显著。其能够识别领域特有词汇，诸如“支持向量机”“混沌时间序列”等词

汇,也能对“海南职业技术学院”这样的机构名命名实体进行良好的识别。

随后以人手工分词 100 篇摘要为标准集,与引入领域词典和未引入领域词典的 jieba 分词做对比试验。实验结果表明,未引入领域词典的 jieba 分词的平均准确率为 87.42%,引入了领域词典的 jieba 分词的平均准确率为 97.44%。领域词典的引入,使得分词的平均准确率上升了 10%。

5.3 学术资源词向量的训练

本文采用 Python 编程语言机器学习包 gensim 中的 word2vec 模块来训练查询扩展所需要的词向量。首先利用 5.2 生成的领域词典对 5.1 获取的图情档领域语料进行 jieba 分词、去停用词等操作,形成一个经过处理好的语料库;然后,将语料库加载到 word2vec 算法中,进行词向量的训练,得到一个词向量库,并将该词向量库以二进制的形式保存,方便后面程序的调用。词向量训练时的参数设置见表 6,其中主要参数词向量窗口设置为 5 个词语窗口,词向量维度设置为 200 维。在词向量库的更新方面,将定期载入新的语料,并对其进行训练后同样保存为二进制词向量库,用最新的词向量库替换以往旧的词向量库。

表 6 word2vec 词向量训练参数设置

Word2vec 训练参数设置	参数含义	Word2vec 训练参数	参数含义
sg = 1	0,对应 CBOW 算法;1 则采用 skip-gram 算法	size = 200	词向量的维度,大的 size 需要更多的训练数据,但是效果会更好
window = 5	表示当前词与预测词在一个句子中的最大距离	alpha = 0.025	模型的学习率
min_count = 5	词频少于 min_count 次数的单词会被丢弃掉	hs = 0	1 采用 hierarchica · softmax,0 则 negative sampling(负采样)
iter = 5	迭代次数	batch_words = 10000	每一批的传递给线程的单词的数量

5.4 系统查全率、查准率和 F1 值表现情况

一般而言,信息检索的查全率(recall)和查准率(precision)是人们最常用到的信息检索评价指标。所谓的查全率,是指检索结果集中,相关的结果与应该被检索到的相关结果的比值;查准率是指在检索结果集中,相关的结果与全部检索结果集的比值。

(1) 查全率公式如式(4)所示:

$$\text{Recall} = \frac{\text{被检出的相关结果}}{\text{全部相关结果}} \qquad 式(4)$$

(2) 查准率公式如式(5)所示:

$$\text{Precision} = \frac{\text{被检出的相关结果}}{\text{全部被检出的结果集}} \qquad 式(5)$$

(3) F1 值。在检索过程中,人们一般想要同时获得最高的查全率和查准率。但是在实际操作中往往达不到如此的效果,一个极端的情况是,如果只返回一个正确的检索结果,那么其查准率会是 100%,而此时的查全率却极低;或者系统耗费巨大的代价将全部相关文献找回,但是此时往往也会带回大量的无关文献,导致查准率极低。需要一种评价方案,能够同时考虑查全率查准率,对检索系统进行评估。

F1 值就是一个很好的选择,能够综合考量两者的影响,其计算方式如式(6)所示:

$$F_1 = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Reecision}} \qquad 式(6)$$

因此,本文还选取 F1 值对检索系统性能进行评估,以从整体对检索系统进行评估。

分别选取与“档案文化”“数据挖掘”“信息素养”这 3 个关键词相关的文献元数据 50 篇,和与这 3 者都不相关的文献元数据 150 篇,总计 300 篇文献元数据进行检索性能测评,测评选取的指标为上文提到的查全率、查准率和 F1 值,结果如表 7 所示:

表 7 系统查全率和查准率

检索词	使用方法	检索结果	相关结果	查全率	查准率
档案文化	基于关键词	40	26	52.00%	65.00%
	本文方法	53	35	70.00%	66.04%
数据挖掘	基于关键词	46	31	62.00%	67.39%
	本文方法	55	38	76.00%	69.09%
信息素养	基于关键词	39	28	56.00%	71.79%
	本文方法	53	41	82.00%	77.36%

注:本文方法是指基于词向量的语义查询扩展方法

将上述查全率和查准率结果转化为柱形图,以直观地展现系统表现情况,详见图 5 与图 6:

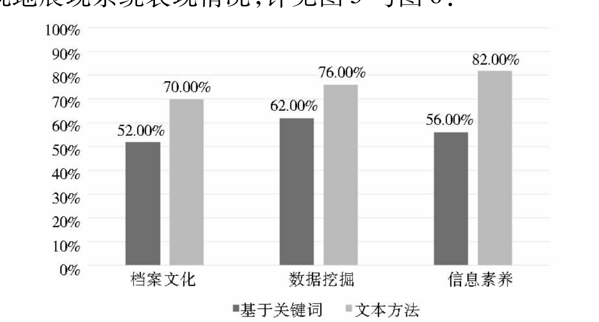


图 5 系统查全率

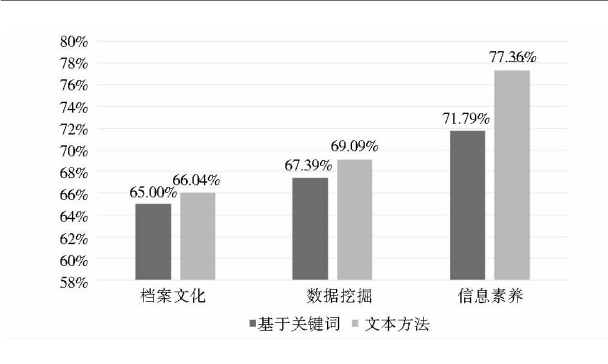


图 6 系统查准率

从表 7、图 5、图 6 可以直观地感受到，与以往的基于关键词匹配的查询方法相比，利用词向量进行语义查询扩展的检索方法在查全率的提升方面要较查准率的提升更为显著：在查准率方面，基于词向量来进行语义查询扩展的检索方法的检索效果平均要比基于关键词匹配的检索方法的检索效果要高 2.77%；而在查全率方面，基于词向量来进行语义查询扩展的检索方法的检索效果平均要比基于关键词匹配的检索方法的检索效果要高 19.33%。造成这种现象的原因也不难解释：与传统的基于关键词匹配的检索方法相比，基于词向量来进行语义查询扩展的检索方法能够将更多相关的检索词加入到检索活动中去，因此能够带回更多相关的检索结果，表现为查全率的显著提升；但是在依靠扩展出来的查询词提升查全率时，其又会带来很多无关的但是也包含相应检索词汇的文档，系统查准率压力较大。需要依靠提升扩展词选取阈值、改进检索相关性算法等手段，控制检索结果，表现为查准率的轻微提升。

最后，我们将系统的查全率和查准率综合考虑，以 F1 值考察系统查全率、查准率，结果如图 7 所示：

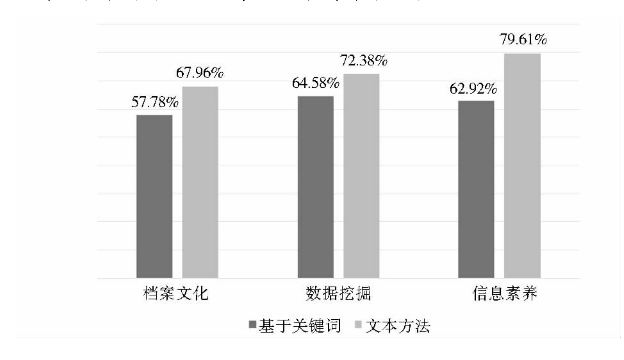


图 7 系统 F1 值

从图 7 中可以看出，系统在 F1 值上也有所提升，几次查询下来，其 F1 值平均提升了 11.56%。

从系统查全率、查准率和 F1 值的整体表现，可以了解到，现阶段应用词向量语义扩展查询的检索方法，较之以往的关键词匹配方法要有优越性，其能够提升系统检索的效果，尤其是系统的查全效果。

6 结语

本文以 word2vec 为词向量生成工具，以 Elasticsearch 为全文检索引擎搭建语义检索系统并针对其中的关键技术——领域词典自动化构建技术、词向量语义扩展技术和针对学术文献的个性化评分方案进行设计。对采集到的图情档领域 12 万余篇学术文献的语义检索实验表明，其能够明显提升信息检索效果，对今后语义检索研究有一定借鉴意义。

本研究对于语义检索结果的排序算法以及以后用于个性化的学术文献推荐还没有展开，有待进一步研究。

参考文献：

[1] 王洁慧. 高校科研用户对图书馆一站式资源发现平台的功能需求分析[J]. 情报理论与实践,2014(12):95－98,80.
[2] FURNAS G W. The vocabulary problem in human-system communication [J]. Communications of the ACM, 1987, 30(11):964－971.
[3] MOLDOVAN D I, MIHALCEA R. Using wordnet and lexical operators to improve Internet searches[J]. IEEE Internet computing, 2000, 4 (1): 34－43.
[4] 高雪霞,炎士涛. 基于 WordNet 词义消歧的语义检索研究[J]. 湘潭大学自然科学学报,2017(2):118－121.
[5] 王李冬,张慧熙. 基于 HowNet 的微博文本语义检索研究[J]. 情报科学,2016(9):134－137.
[6] BLEI D M, Ng A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of machine learning research, 2003(3):993－1022.
[7] BLEI D M, LAFFERTY J D. Correction: a correlated topic model of science[J]. Statistics, 2007, 1(1):17－35.
[8] 刘启华. 基于 LDA 的文本语义检索模型[J]. 情报科学, 2014 (8): 38－43,55.
[9] GOOGLE. Word2vec [EB/OL]. [2017－08－26]. https://code. google. com/archive/p/word2vec/.
[10] MIKOLOV T. Word2vec [EB/OL]. [2017－08－26]. https://github. com/tmikolov/word2vec.
[11] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2018－06－17]. https://arxiv. org/pdf/1301.3781v3. pdf.
[12] 范桥青, 方钰. 面向健康问答社区的语义检索技术研究与分析[J]. 电子技术与软件工程, 2017(2):202－204.
[13] 刘梦兰, 刘斌, 彭智勇. 基于词向量的专利自动扩展查询研究[J]. 计算机工程与科学, 2017(12):2297－2305.
[14] 许稳堂. 基于词向量的微博检索系统研究与实现[D]. 上海: 东华大学, 2017.
[15] STANFORD. GLOVE [EB/OL]. [2017－08－26]. https://nlp.

stanford. edu/projects/glove/.
[16] 陈国华，汤庸，许玉赢，等. 基于词向量的学术语义搜索研究[J]. 华南师范大学学报(自然科学版)，2016，48(3):53－58.
[17] GOOGLE [EB/OL]. [2018－06－17]. https://research.google.com/semanticexperiences/.
[18] 张榕. 术语定义抽取、聚类与术语识别研究[D]. 北京:北京语言大学，2006.
[19] ELASTICSEARCH [EB/OL]. [2017－08－26]. https://www.elastic.co/cn/.
[20] 张孝飞，孔繁秀. 基于语义概念分析的科技文献检索研究[J]. 情报理论与实践，2016，39(8):115－118.
[21] PAGE L，BRIN S，MOTWANI R，et al. The pagerank cita-tion ranking: bringing order to the Web [R]. Stanford InfoLab，1999.
[22] 百度百科. 2017 年最新版《中文核心期刊要目总览》[EB/OL]. [2017－08－26]. https://wenku.baidu.com/view/15c20df10d22590102020740be1e650e52eacfa4.html.

作者贡献说明：

王仁武：负责论文选题与框架确定，实验设计，论文修改和审阅；

陈川宝：负责开展实验，论文初稿撰写；

孟现茹：负责数据采集与处理，开展实验。

Semantic Retrieval Technology of Academic Resources Based on Word Embedding Extension

Wang Renwu　Chen Chuanbao　Meng Xianru

Department of Information Management, Faculty of Economics and Management,

East China Normal University, Shanghai 200241

Abstract: [**Purpose/significance**] Based on the statistical method, the paper explored the semantic retrieval technology based on word embedding expansion to enhance the semantic retrieval ability of academic resources. [**Method/process**] Using Natural Language Processing and text mining technology, the paper preprocessed the collected academic resources (mainly academic papers) metadata, combined the Word2vec word embedding generation tool and the elasticsearch full text retrieval engine to build semantic retrieval system, and explored the semantic retrieval of academic resources. [**Result/conclusion**] The method proposed in this paper can effectively improve the retrieval effect of academic information, and it realizes the semantic retrieval of academic resources to a certain extent, and could provide reference for further research on the follow-up semantic retrieval.

Keywords: Word2vec　elasticsearch　semantic retrieval　academic resources